Fraunhofer SCAI

SCAIView 1.0

(Human Version)

SCAIVIEW

The Knowledge Discovery Framework

User Manual

Edited by:

Dr. Christoph Friedrich, Erfan Younesi

Last Update: May 2010

Disclaimer

This system is provided by the Fraunhofer SCAI “as is" without warranty of any kind. We may modify or halt this system at any time without prior notification. We do not warrant or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed.

This system is built on the Medline database leased from the National Library of Medicine [NLM]. Title and MeSH Headings are adopted from MEDLINE®/PubMed®, a database of the U.S. National Library of Medicine.

We cannot assume any liability for the content of external pages. Solely the operators of those linked pages are responsible for their content.

We make every reasonable effort to ensure that the content of this Web site is kept up to date, and that it is accurate and complete. Nevertheless, the possibility of errors cannot be entirely ruled out. We do not give any warranty in respect of the timeliness, accuracy or completeness of material published on this Web site, and disclaim all liability for (material or non-material) loss or damage incurred by third parties arising from the use of content obtained from the Web site.

Registered trademarks and proprietary names, and copyrighted text and images, are not generally indicated as such on our Web pages. But the absence of such indications in no way implies that these names, images or text belong to the public domain in the context of trademark or copyright law.

System Requirements

- Firefox Browser >2.0.x.x, Safari or Internet Explorer> 6.0, Google Chrome, and Opera
- 1 GB of RAM and a hardware generation >=2005
- Username/Password required and can be obtained by E-mail from: friedrich@scai.fraunhofer.de

Table of Contents

1 INTRODUCTION 1

1.1 Development 1

1.2 License 1

2 QUICK START 2

3 DETAILED EXPLANATION 3

3.1 Search Component 3
3.1.1 Button Description 3
3.1.2 Search Field 3

3.2 Search Examples and Explanations on queries 4

3.3 Entity Tree Component 7
3.3.1 Entity Classes 7
3.3.1.1 Tree View 13
3.3.1.2 Button Description 13

3.4 Result Component 14
3.4.1 Entity View 14
3.4.1.1 Entity Tab Components 16
3.4.1.2 Entity Table Columns 16
3.4.2 Document View 25
3.4.3 Analysis View 27

3.5 Application Scenarios 28

1 Introduction

SCAIView is an advanced semantic search engine that addresses questions of interest to general biomedical and life science researchers. Most of the current knowledge exists as unstructured text (publications, text fields in databases) and SCAIView provides users with full text and biomedical concept searches which are supported by large biomedical terminologies and outstanding text mining technologies.Using machine learning and dictionary-based Named Entity Recognition (NER), SCAIView extracts information of genes, drugs, SNPs and other Life Science entities from MEDLINE abstracts. SCAIView uses a multi-threaded Lucene to allow semantic and ontological search on this data. Documents are retrieved via free-text queries chosen by the user and a span of biomedical entities such as genes/proteins, SNPs, drugs, etc. can be selected from the terminologies and ontologies. Complex queries can be asked such as “what drugs are mentioned in the context of Alzheimers disease”? or “what genes are co-mentioned with Diabetes and are on the insulin signalling pathway”?

1.1 Development

SCAIView has been developed and maintained by the bioinformatics team of the Fraunhofer Institute for Algorithms and Scientific Computing, SCAI. The selected biomedical entities are found by an approximate search algorithm implemented in the Fraunhofer-Gesellschaft information extraction tool, ProMiner®, which additionally disambiguates synonyms of entities to unique identifiers in public available databases.

Visit www.scai.fraunhofer.de/scaiview.html?L=1 for more information.

It must be noted that the development phase of this system has been partially funded by @neurist project in the framework of the European integrated project.

1.2 License

SCAIView-Human Version is free for academic use; commercial users and those users who wish to access the API for large queries must contact Dr. Christoph M. Friedrich via friedrich@scai.fraunhofer.de.

The content of our database might be accessed for copying purposes but we do not allow bulk downloads. SCAIVIew-Human version also includes a number of other open source libraries, which are detailed in the User Manual/Acknowledgements below.

2 Quick Start

Step 1- In the grey search field, type your query. The query could be a disease name, a biological process, the title of a journal, the name of an author, or the PubMed identification number of an article. A number of predefined query terms are provided in the dropdown menu.

Step 2- Select from the entity tree what you are looking for in your query results: genes/proteins, SNPs, Chromosomal locations, GO annotations, MiRNAs, etc. Click on the entity class of interest in the tree only once and make sure that your selection turns into green with the magnifier in front of it. Leave the confidence level on the default (the level 5 returns the most stringent results).

NOTE: Clicking any entity class twice turns it to the red colour meaning that this class is "excluded" from the search.

Step 3- Press search button. The page is redirected to the Entity tab where the results are listed and ranked according to the relative entropy score. By default, 10 entities per page are shown and the user can navigate between the pages but it is also possible to see more entities per page by selecting through the dropdown menu to the right side of the page navigation.

Step 4- Click on one of the entities of your interest from the result list. You will be directed to the Document tab where all PubMed abstracts that contain this entity and are relevant to the query are shown. To see the frequencies of all entities over all found abstracts, return to the Entity tab and click on the analysis icon. You will be directed to the Analysis tab where an overview of the entities and their occurrence frequencies in the abstracts is given.

Step 5- In the Document tab, you are able to highlight your entities of interest in the text by selecting the colour-coded sections at the top of the page. PMIDs or comments can be exported to a text file.

Step 6- To start a new search click.

Further selection and filtering of results are possible by using search component as described below.

3 Detailed Explanation

3.1 Search Component

On the top-left side of the user interface the Search Component is located with several buttons and a search field.

3.1.1 Button Description

- Decrease the font size
- Increase the font size
- Reset the Search
- Filter the Results
- Show the Information Screen
- Start the Search
- Select from predefined queries

3.1.2 Search Field

In the grey search field, located below the icons, you can either enter a string, e.g. a disease name, or select from predefined queries by clicking the blue down arrow button right to the search field.

3.2 Search Examples and Explanations on queries

In the search field you have the possibility to use certain keywords to make your search more specific. It works like any other search engine, but knowing the special features allows you to be more effective with your queries and allows the proper interpretation of the results:

1) The boolean function AND is automatically considered between multiple keywords except that the user indicates otherwise.

2) Performing null query (empty search field) on any entity class results in retrieval of entire entities for that class from PubMed abstracts.

3) It is also possible to define a subcorpus from PubMed database by invoking the E-utilities (or programming utilities) from inside SCAIView. For instance, if you want to analyze the abstracts that you have retrieved in PubMed database, you can type “EUTILS{…your query term...} in the search field and obtain the entity analysis results on this set of abstracts.

We explain the search possibilities by the following examples:

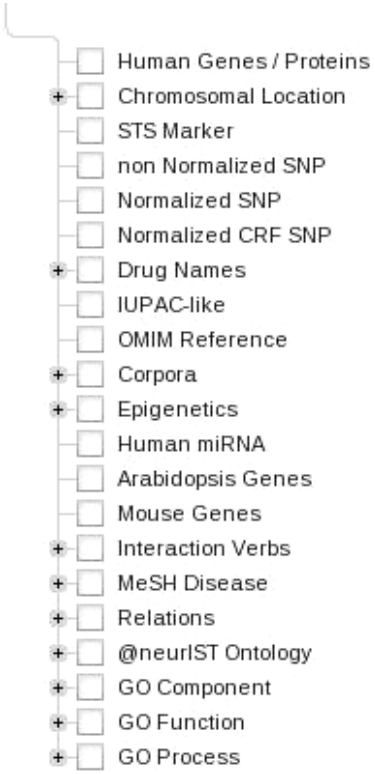
Search query	Description
Inflammation	**Occurrence:** Find all documents containing the word ‘inflammation' in any of the main text fields (title, abstract, PMID or MeSH). It will find all occurrences of versions of the word ‘inflammation’, even if they contain capital letters (case insensitive).
PMID:(19551867 OR 19833996)	**Document Identifier:** Find the documents with the Medline Identifiers (PMID) 19551867 or 19833996.
inflammation AND stroke	**Conjunction:** Find all documents containing both the word 'inflammation' and the word 'stroke' in any of the main text fields (title, abstract or MeSH). The two words may be in different fields. The operator AND has to be in capital letters.
proinflammatory OR antiinflammatory	**Disjunction formal:** this form of query finds all documents containing the word “proinflammatory” or “antiinflammatory”. The Operator OR has to be in capital letters.

h*moglobin production	**Wildcard *:** Find all documents containing the word 'production' in any of the search fields, and the word 'hemoglobin' or ‘haemoglobin’ or ‘hxxxxmoglobin’. The asterisk is used as a possible replacement for any subtext.
carcinoma AND DATE:[1980 TO 2100]	**Date Range:** Find all documents containing the word 'carcinoma' in any text field which have a publication date between the year 1980 and the year 2100 (the 2100 is a replacement for up to the newest). This is quite useful to avoid, false positive hits, like the Gene ‘AIR’, which is often found in the old Medline entries prior to 1975 (the titles are fully capitalized).
anaemia AND MESH:genetics	**MeSH search:** Similar to the previous query, this search query avoids false positive matches. It does this by restricting the search only to the documents that have been assigned to the MeSH (Medical Subject Headings) category of genetics in addition to the term ‘anaemia’. Please note that the human MeSH annotators at NCBI are slower than the publications. This means that it takes up to 2 years to fully categorize the publications; in the meantime you will not find them with these restrictions.
(proinflammatory OR inflammation) AND (human OR mouse)	**Groupings:** Find all documents that contain either the word 'proinflammatory' or the word 'inflammation', and which also contain either the word 'human' or the word 'mouse'. Note that without the parenthesis this query would be interpreted in an entirely different manner. The operators AND/OR have to be in capital letters.

"breast cancer"~5	**Spanned Search:** Find all documents containing the word 'breast' within 5 words of 'cancer', in any of the text fields. The 5 may be replaced with any integer.
chromosome*	**Wildcard *:** Find all documents containing words starting with the prefix 'chromosome' like 'chromosomes', 'chromosomal', etc (See the Note).
su?	**Wildcard ?:** Find all documents containing words that have only one character after 'su' like sub, sun, sum etc.
AUTHORS:Hofmann	**Author:** This finds all documents, where 'Hofmann' occurs as a co-author.
JOURNAL:Stroke	**Journal search:** Finds all documents, where the Journal name contains Stroke.
EUTILS{alzheimer}	**PubMed E-Utilities:** Using this command directly typed into the search field enables the user to pull the relevant/selected abstracts directly from PubMed database into the SCAIView environment for entity recognition analysis.

Note: Caution should be taken in using asterisk wildcard option. Using asterisk wildcard symbol at the end of query keyword enforces the system to apply the "stemming" functionality for finding terms that have endings other than the usual form. For example, querying the system for 'Alzheimer' on the MeSH disease returns more than 65000 documents but using the asterisk wildcard as 'Alzheimer*' returns only 24 documents containing rare variations of the term Alzheimer such as 'Alzheimerization', 'Alzheimer-apoE4' , 'Alzheimerism' etc.

3.3 Entity Tree Component



The entity tree component is used for the selection of the different Entity Classes that are of interest to the user. It includes all classes that are indexed from Medline by our entity recognition tools. If the entity tree component is not fully expanded, a plus '+' sign, shows that a subtree is present. Clicking on the plus expands the subtree and allows choosing from sub-components.

3.3.1 Entity Classes

The tree component allows choosing several different entity classes of interest (singly or in combination).

Note: When switching between different entity classes over the tree, make sure that you've deselected the previous selection except that you intend to perform your query over multiple entity classes simultaneously to get more focused results.

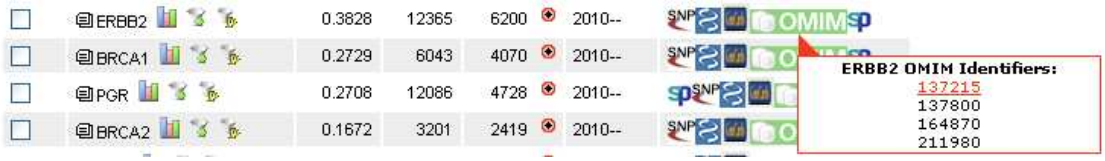
Genes / Proteins:

The entities of the Class 'Genes/Proteins' are found by ProMiner software through an approximate string search and using the dictionary that is generated of synonyms found in the databases EntrezGene and Swissprot and normalized to those IDs. There are four separate Gene/Protein classes in the tree for four organisms: cattle genes, pig genes, mouse genes/proteins, and human genes/proteins. Link-outs of this Entity class are provided to the following external databases:

- EntrezGene at NCBI that provides information on genes;
- HGenetInfoDB developed at IMIM that provides information on SNP of genes;
- GeneCards® is a searchable, integrated database of human genes that provides concise genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes;

- OMIM The Online Mendelian Inheritance in Man at NCBI is a database that catalogues all the known diseases with a genetic component and links them to the relevant genes in the human genome and provides references for further research.
- sp swissprot SwissProt/UniProt that provides information on proteins. A gene might produce several proteins, e.g. splicing variants, so you may find several instantiations of this icon for a single entry.

Note: *For some entities there may exist several entries in the same database; for example, a gene or protein may have several entries in the OMIM database. Mouse-over or click on the linkout icons presents a list of identifier numbers for these entries and another click on each identifier number redirects you to that entry page in the OMIM database.*



Chromosomal Location:

The entities of the Entity Class 'Chromosomal Location' are found by a regular expression searching for Cytoband information, done with ProMiner. This information is frequently used in Linkage Analysis and a search gives an overview on the involvement of genetic information on Chromosomes in relation to a query.

STS Marker:

The entities of the Entity Class 'STS Markers' are found by a regular expression search of the Identifiers, executed by ProMiner. STS Markers are not only used in Linkage Analysis but also used to relate sequential information to clearly defined positions on the sequence. Later versions will include link-outs to the UniSTS database at NCBI: STS

non-Normalized SNP:

The entities of the Entity Class 'non-normalized SNP' are found by MutationFinder[1] system. They consist of mutation mentions from text or Variation mentions compliant with the Mutation Nomenclature[2] found by the regular Expression facility of ProMiner[3].

[1] Caporaso JG, Baumgartner WA Jr, Randolph DA, Cohen KB, Hunter L, Mutation-Finder: A high-performance system for extracting point mutation mentions from text, *Bioinformatics* 23(14):1862–1865, 2007

Normalized SNP:

The entities of the Entity Class 'Normalized SNPs' are found through a search performed by OSIRIS[4], while the dictionary is generated from the synonyms found in the EntrezSNP database and normalized to those IDs. Direct mentions of dbSNP identifiers are found by the regular expression feature of ProMiner with inclusion and exclusion criteria.

Link-outs of this entity class are provided to:

- , NCBI EntrezGene at NCBI, provides information on genes.
- SNP SNP HGenetInfoDB developed at IMIM, provides information on SNP of genes.
- , GeneCards® GeneCards provides concise genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes;
- dbSNP dbSNP at NCBI, provides information on genetic Variations.
- HapMap HapMap describes the common genetic variants in human genome.

Note: A search under the Entity Class "Normalized SNP" will highlight SNP co-mentionings in the Document View. Hovering your mouse over the highlighted SNP leads to a link-out menu pop-up containing a header line about the corresponding dbSNP code, type of mutation, chromosomal location, gene name, and the array platform (Affymetrix or Illumina or both) as well as links to the Entrez gene database, HGenetInfoDB database, GeneCards database, dbSNP and HapMap databases. Further information on the array platforms can be obtained by following the link-outs. In some cases it can be seen that a SNP occurs within the sequences of two genes simultaneously. Such information is included in the header line describing the name of both genes as well as additional link-outs for both genes to their corresponding entries in Entrez gene, SNP, and GeneCard databases.

---

[2] den Dunnen, J. T. & Antonarakis, S. E. Nomenclature for the description of human sequence variations. *Hum Genet,* 2001*, 109*, 121-124

[3] Roman Klinger; Laura I. Furlong; Christoph M. Friedrich; Heinz Theodor Mevissen; Juliane Fluck; Ferran Sanz & Martin Hofmann-Apitius, Identifying Gene Specific Variants in Biomedical Text *Journal of Bioinformatics and Computational Biology,* 2007*, 5*, 1277-1296

[4] Bonis J, LI Furlong, F Sanz. OSIRIS: a tool for retrieving literature about sequence variants. Bioinformatics 22:2667-2569 (2006).

Normalized CRF SNP:

The entities of this class are the SNP mentionings in the text that are found by CRF (Conditional random Field) algorithm. CRF is a machine-learning method which is best suited for sequential data.

Drug Names:

The entities of the Entity Class 'Drug Names' are found by an approximate string search performed by ProMiner while the dictionary is generated of the synonyms found in the Drugbank[5] database version 2 and normalized to those IDs. Drugbank provides information to more than 4000 different drugs and link-out is provided via: DrugBank.

IUPAC-like:

The entities of this class consist of the names of chemical entities which follow the standard naming rules based on IUPAC nomenclature. These entities are extracted from text using a new machine learning approach based on conditional random fields[6].

OMIM Reference:

The entities of the Entity Class 'OMIM (Online Mendelian Inheritance in Man) References' are found by a regular expression search for the IDs of the OMIM database, performed by ProMiner. Link-outs are provided to OMIM at NCBI.

Reference corpora:

Under this class, 7 subclasses are embedded which contain collections of structured literature texts from specific resources (publications dealing with the topics related to the Alzheimer's, Parkinson's, and Schizophrenia) as well as general full-text publications from PubMed Central database. The subclass "Full text" makes it possible to analyze those abstracts that are exclusively found in PubMed Central repository with access to the corresponding full texts. By selecting the Fulltexts (ftp) subclass, it will be possible to access and download the full-text

[5] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. & Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res, Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8. david.wishart@ualberta.ca,* 2006, 34, D668-D672

[6] Klinger, R.; Kolárik, C.; Fluck, J.; Hofmann-Apitius, M. & Friedrich, C.M. Detection of IUPAC and IUPAC-like chemical names. Bioinformatics, 2008, 24, i268-i276

articles from the PubMed FTP service. The option of Systematic Review enables the user to see the results of text analysis on the abstracts of systematic reviews from PubMed.

Epigenetics:

Under this entity class, subclasses of histone modification are assigned. Using this option, it is possible to detect histone modifications in biomedical literature with Conditional Random Fields[7].

Human miRNA:

This entity class enables the user to find microRNA named entities in the text with the possibility of access to the miRBase database[8] through linkouts from both the Entity view page as well as the annotated text.

Arabidopsis Genes:

Selecting this entity class highlights the gene names specific to the plant model organism, *Arabidopsis thaliana*, in the abstract texts.

Mouse Genes/Proteins:

This entity class contains a collection of gene and protein names specific to the model organism, *Mus musculus* and once selected, the relevant gene names and corresponding synonyms are identified in the abstract texts.

Interaction Verbs:

This option highlights the type of interactions mentioned in the text. These are interaction verbs which are mentioned in a biologically-meaningful context in the text.

MeSH Disease:

This Entity Class contains all the Disease names that exist as Medical Subject Headings (MeSH). Inclusion of this option in the search allows the coverage of all disease aliases related to the query keywords.

---

[7] Kolářik C., R. Klinger, and M. Hofmann-Apitius. Identification of Histone Modifications in Biomedical Text for Supporting Epigenomic Research. BMC Bioinformatics, 10(S28), January 2009.

[8] Griffiths-Jones S, HK Saini, S van Dongen, AJ Enright. miRBase: tools for microRNA genomics. NAR 2008 36(Database Issue): D154-D158.

Relations:

Relations describe the general associations found between the entity of question and other nearby entities in the text. For example, if we are searching for the genes/proteins related to the query term 'breast AND cancer' which additionally have a certain positive or negative association to a disease or drug, we can determine the type of association in Relations subtree. In this case, the found expression may look like this: "Forkhead box A1 expression in breast cancer *is associated with* luminal subtype" or "the F31I polymorphism in AURKA *is not associated* with a modified risk of breast cancer in BRCA1 and BRCA2 carriers".

These associations can be further specified for finding positive/negative associations between the entity in question and a gene or a SNP, specifically. Controversial associations are those associations which are found to be contradictory in the scientific literature.

@neurIST Ontology:

The entities of the Entity Class '@neurIST Ontology' are found by an approximate string search with ProMiner while the dictionary is generated from the 'particular in text-mining' part of the @neurIST Ontology and normalized to those IDs. Link-out is provided to the Ontology Browser with definitions at UKFLR (username/password needed) via: @neurIST

The ontology comprises of further sub-trees of aneurysm disease terminology which can be used to narrow down the search in the corpus. The ontology covers aneurysm specific clinical terms such as diagnostics, therapy, and risk factors.

GO Component:

This class represents the biological cell compartments that are defined by Gene Ontology and contains further subclassifications for detailed search.

GO Function:

This class contains subclassifications of Gene Ontology for gene biological function and can be used for supporting the query for finding information attributed to the gene at functional level.

GO Process:

This class covers the subclassifications of those terms in Gene Ontology that describe the biological processes for each gene.

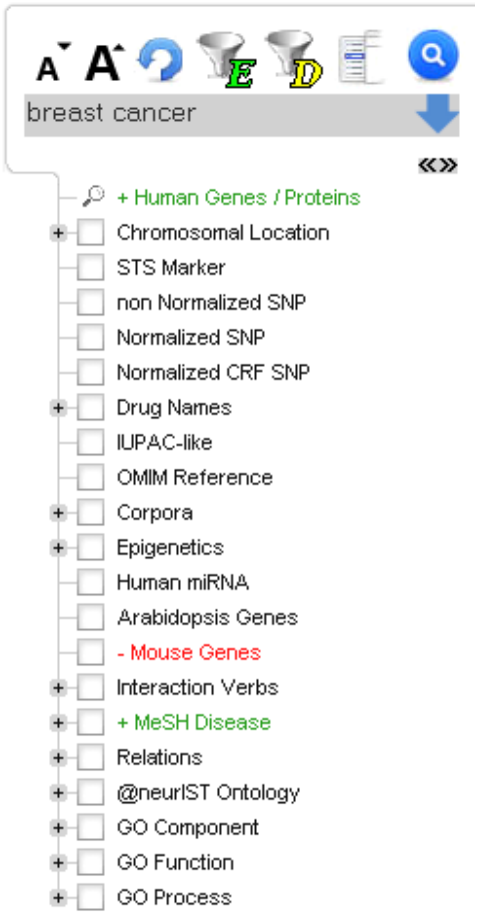
Confidence levels:

By choosing levels of confidence ranging from 1 to 5, it is possible to adjust the level of accuracy in the results. In the Document view, mouse-over on entities shows the corresponding level of confidence as well.

3.3.1.1 Tree View

The tree view provides several different selection types:

- Click once on the name of an item in the tree **to include it** into the search (a little **plus** is shown)
- Click once again to **exclude it** (a little **minus** is shown) and
- Click again to **disregard** it.



3.3.1.2 Button Description

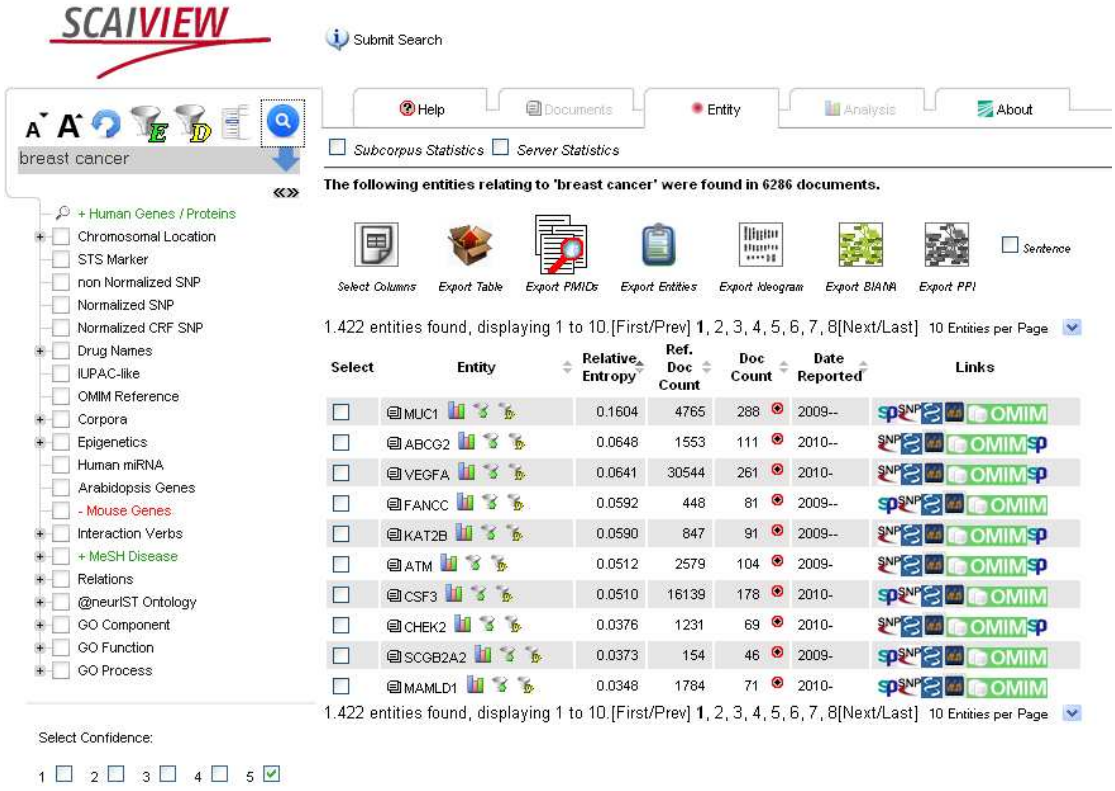
- «» Expand / Collapse Tree Viewing
- Show this Entity Class in the Entity View

3.4 Result Component

It displays the results according to the searches and can be navigated via tabs. You find an Entity View, an Analysis View and a Documents View. In the “Entity View”, named entities of interest are summarized under the column “Entity” and are directly linked to their corresponding abstracts. By rolling mouse over each named entity under Entity column, the full name of the entity (e.g. full gene name) as well as its identifier will be shown. Shifting to the analysis tab takes the user to the “Analysis View” where the collective information about the type and number of all other entities that co-occurred with the entity of interest in the text can be found.

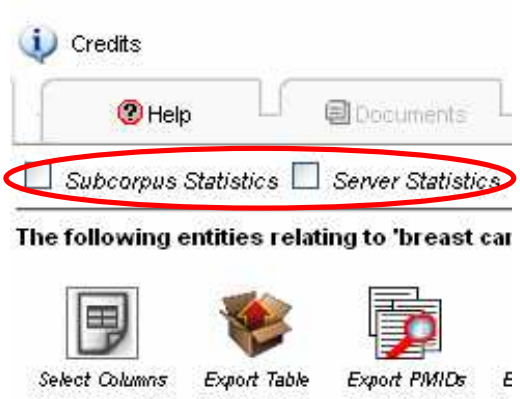
3.4.1 Entity View

The Entity View displays a table with the aggregated list of entities found in the documents ranked by a certain column. Clicking on a column-Header of the table will re-order the table depending on the selected column entries. You may have to click twice to switch the order. You can navigate between pages either by clicking Next/Last or switching directly to a specific result page by clicking on the page number. Using the dropdown menu In front of the page navigator, it is possible to determine the desired number of entities to be shown on the page.



Statistics:

To obtain detailed information on the conducted query, you can expand two statistics tables at the top of the entity view. By ticking the “Server Statistics” you obtain information on the performance of SCAIVIEW system during information retrieval for your query whereas “Subcorpus Statistics” provides a quick overview of all entities found in the subcorpus of your query (see below).



By clicking on each entity class, an overview table appears containing information on the individual named entities and their corresponding relative entropy, document count, and link-outs, if any. The content of these tables can be exported to CSV, Excel, and XML file formats via the links provided at the bottom of each table.

Subcorpus Statistics Server Statistics

Chromosomal Locations	Drug Names	Protein/Gene	STS Marker	OMIM Reference	@neurIST	non Normalized SNP	Normalized SNP	MeSH Disease

1.800 entities found

Entity	Relative Entropy	Doc Count	Links
Mammary Carcinomas, Human	3.8064	49699	
Neoplasms	2.4013	31353	
Metastases	0.6381	8331	
Undifferentiated Carcinomas	0.4607	6015	
Cancers, Ovary	0.2922	3815	
Lymphatic Metastasis	0.2183	2850	
Cribriform Carcinomas	0.2058	2687	
PROSTATIC NEOPL	0.1935	2527	
Breast Invasive Ductal Carcinoma	0.1935	2526	
Neoplasm Invasiveness	0.1861	2430	

1.800 entities found

Export options: CSV | Excel | XML

3.4.1.1 Entity Tab Components

***Select Columns*:**

In the Entity table, the following columns are shown by default: Entity, Relative Entropy, Reference Documents Count, Documents Count, and Link-outs. However, the results table can be expanded by adding more columns to include additional information such as links to KEGG or Reactome pathways as well as Gene Ontology, Cytoband, synonym information, SciMago score, InterPro family and domain information, ATC (Anatomical Therapeutical Chemical Classification) code, and HUGO standard gene naming. Moreover, it is possible to indicate whether a protein has been already targeted by any drug and to visualize and download the structural images of the drugs. These options can be added to the result table in the form of columns by ticking them in the drop-down menu. Information on these columns are detailed as follow.

3.4.1.2 Entity Table Columns

- The column '**Entity**' displays the official entity name. If you click on the entity name the Result Component switches to the Document View and all documents shown include the entity and the corresponding search query. You can come back to the entity view by selecting the Tab 'Entity' at the top.

Note: In front of each entity name, there are two icons: the analysis icon which directs the user to the statistics page, and the filter icon that enables the user to copy the selected entities in the Entity view to the filter field so that a secondary search can be performed but restricted to the entities copied in the filter field.

- The column '**Relative Entropy**' displays the relevance of the entity for a given search query. This measure defines the distance of the entity in the corpus of the search string (specific) relative to the complete Medline (completely unspecific). The range of values of relative entropy spans from -1 to 1. The closer this values to 1, the more relevant the entity to the query. The value should be used for comparison purposes only. The approach chosen in the current version uses document frequencies to calculate the ranking, meaning once a document contains the entity in question, regardless of how many times it appears, the document is counted once.

-The column **'Aneurysm Linker Degree'** lists the number of the interacting protein partners in the context of intracranial aneurysm protein interaction network (for more details, refer to the @neurIST project website.

- The column '**Odds Ratio'** ranks the results according to their likelihood of occurrence (alternative to relative entropy).

- The column '**Document Count**' lists the number of documents in the corpus of the search query containing the entity. This corpus is called hit list. Click on the numbers or red circles results in the export of all PubMed identifiers which refer to a specific entity.

- The column '**Full Corpus Document Count**' lists the number of overall documents in the MEDLINE containing the entity.

- The column '**Entity Count**' includes the number of corresponding entity found in the MEDLINE abstracts.

- The column '**Synonyms**' lists all synonym names (aliases) of the corresponding entity (gene/protein or drugs or disease).

- The column **'Structure Image'** shows a thumbnail of the chemical structures for retrieved drug names (source: DrugBank) and clicking on the thumbnail downloads the chemical structure in SDF format *(NOTE: This option only applies to the drug named entities)*.

- The column '**Cytoband**' includes the chromosomal location of the corresponding gene entity in human genome.

- The column '**InterPro family**' not only exhibits the family class to which the protein or protein product of the gene belongs, but also provides a direct link-out to the InterPro database.

- The column '**InterPro domain**' not only lists the name and identifiers of the domains present in the protein in question, but also provides a direct link-out to the InterPro databse.

- The column '**SciMago Score**' lists the SciMago index of the journal where the abstract has been published. This index is similar to impact factor indexing which has been introduced by an independent Spanish university.

- The column '**ATC Code**' is only valid for Drug Names selection and stands for Anatomical Therapeutic Chemical Classification System. This column describes and links the drug name entities to their corresponding classes in the ATC system.
- The column '**Drug target**' determines whether the protein or product of the gene in question is already recognized as a target for specific drugs in the DrugBank database.

- The column '**Entrez Gene Identifier**' can be added to the Entity table from the drop-down menu on top left. It shows the gene identifier for each corresponding gene entity.

- The column '**HUGO**' lists the unique gene symbols for each gene/protein entity, assigned by Human Genome Organization nomenclature committee.

- The column '**KEGG Pathway**' describes the molecular pathway(s) in which the corresponding entity is involved. There is also a link-out to the representation of the pathway in the KEGG database.

- The column '**Reactome Pathway**' describes the biological pathway(s) in which the corresponding entity is found. The link-out redirects the user to the original information in the Reactome database.

- The column '**Gene Ontology**' lists the GO ontology annotations for the corresponding gene/protein entity and each annotation is hyperlinked to the AmiGO definitions.

- The column '**Last Publication Date**' contains the most recent/last publication date that have been found for the documents containing the entity of interest. Reordering on this column gives you a ranking on the entities.

- The column '**Link outs**' provides several links to external databases (e.g SwissProt,, NCBI, HGenetInfoDB, etc.

Export Table:

By clicking this icon, the full results of query search including all the entities found in the literature and their relevant information shown in the result table can be exported to the text file (CSV format). It is possible to export the results for each entity type selected in the subtree (e.g. gene/protein, Drug Names, etc; if applicable).

Export PMIDs:

This functionality allows the user to export the list of all the extracted entities (e.g. gene names) along with their corresponding PubMed identifier and thus provides a means for tracing back the reference from which the entity has been extracted.

Export results to clipboard:

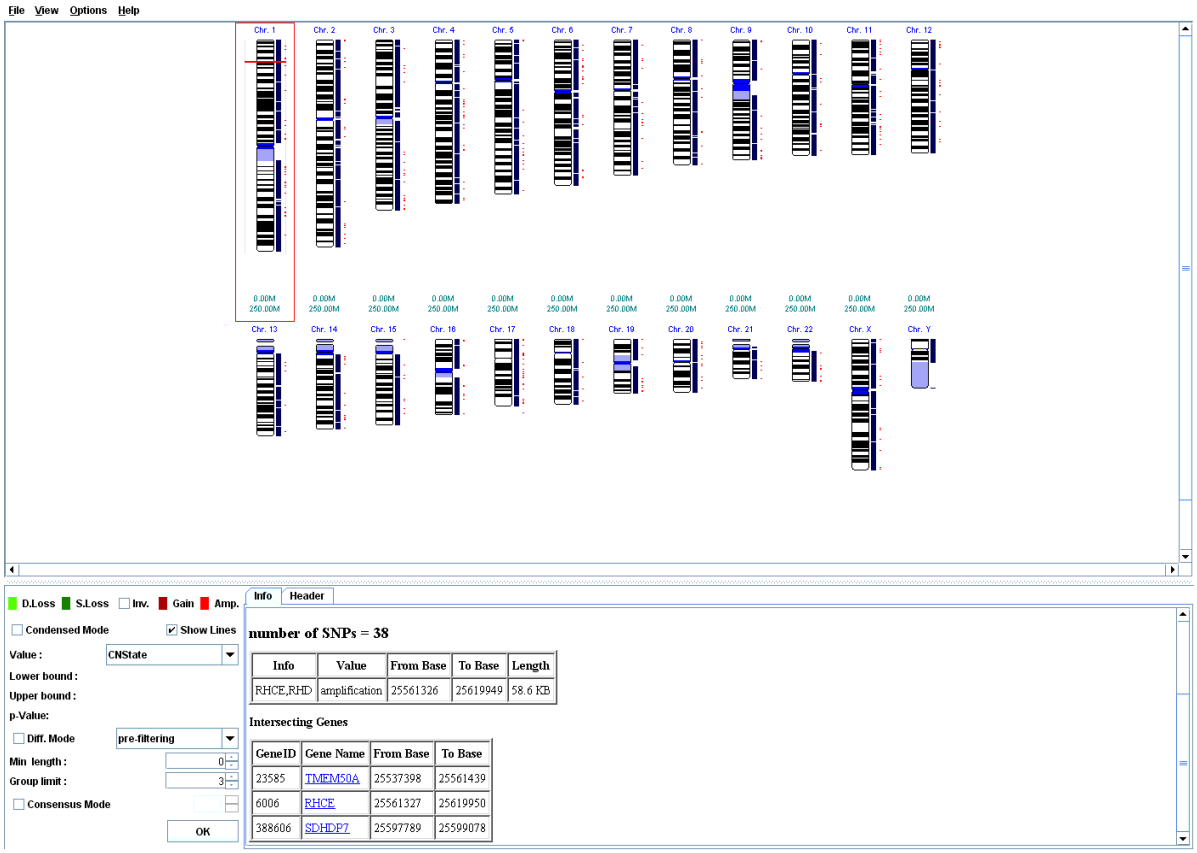
This option allows users to export selected entities from the Entity View to the clipboard. Desired entity names should be check-marked first in the Select column; then a click on the clipboard icon opens a small window containing the selected entities. This functionality allows users to copy the selections or to save them as a text file and use them as direct input for other applications such as GO ontology analysis by BiNGO, a plugin of Cytoscape software.

Export SNP results:

This functionality is most useful for the visualization of literature-extracted SNPs in the human karyogram. Note that you need to select Genes/Proteins in combination with Chromosomal Location in order to get a meaningful output (first select the Chromosomal Location, then select the SNP Normalized, and press search). The output is exported to a text file by clicking on the above icon. For visualization of SNP markers on the human karyogram, you must install the software “Ideogram Browser”[9], freely available for download at www.informatik.uni-ulm.de/ni/staff/HKestler/ideo . The text output file can be loaded into IdeogramBrowser from the Files --> Load Markers. Depending on the type of SNPs (gain vs loss), the imported markers are visualized beside chromosome locations of relevant genes. Clicking on the

[9] Mueller A, Holzmann K, Kestler HA, Visualization of genomic aberrations using Affymetrix SNP arrays, Bioinformatics, 23(4):496-497, 2007

imported marker line beside each chromosome provides more information regarding the SNP numbers and contents in the Info tab. For example, if we search for the SNPs related to '*Intracranial AND aneurysm*' and export the data using the above option, it can be used as input in Ideogram Browser to visualize the coordinations of corresponding SNP(s) on the human chromosomes.



Export protein-protein interaction network from BIANA[10]:

Using this option, it is possible to construct information-enriched protein-protein interaction networks around a set of "seed" protein/gene entities by selecting those entities under the Select column; then clicking on the above icon will generate an interaction file in the XGMML format which can be directly imported into Cytoscape environment for network visualization. The advantage of this format is that it can include additional information about the biological relationships of the network elements. By clicking on the above icon, the user can select

[10] http://sbi.imim.es/web/BIANA.php

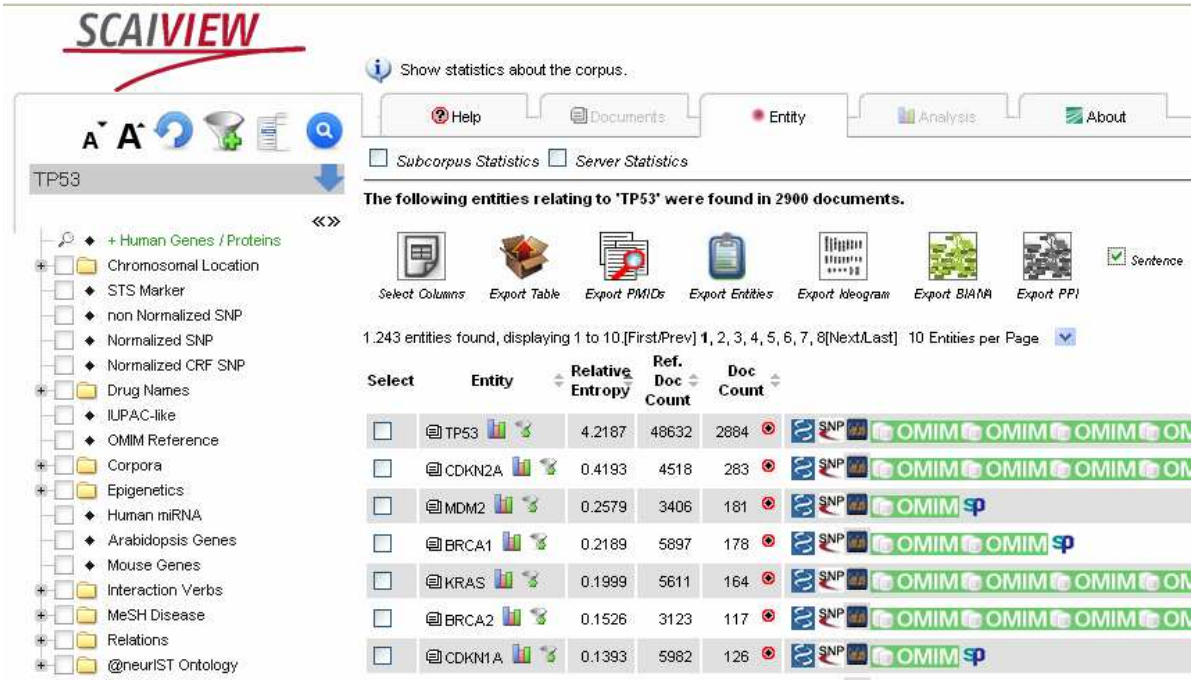
which type of data should be included in the file. The following data types can be included in the network file: Entrez gene identifier, KEGG and Reactome pathway, GO annotations, dbSNP, PFAM, and PDB. This information is embedded into the XGMML file by BIANA automatically and can be visualized in Cytoscape environment under the Data Panel by selecting the attributes.

Export co-citation results:

This option allows users to download the co-citations of each entity type (e.g. gene/proteins or drug names) in the literature. The frequency of co-occurrences found for the entities as well as other informative attributes including the corresponding Entrez gene identifier is shown in separate columns. This option mainly aims at exporting the results of protein-protein co-citations in a tab-delimited text format which enables users to input the file directly into Cytoscape software for visualization as well as topological analysis of the co-citation network. In Cytoscape, the entities are translated into nodes and their co-occurrence frequencies into the attribute of edges.

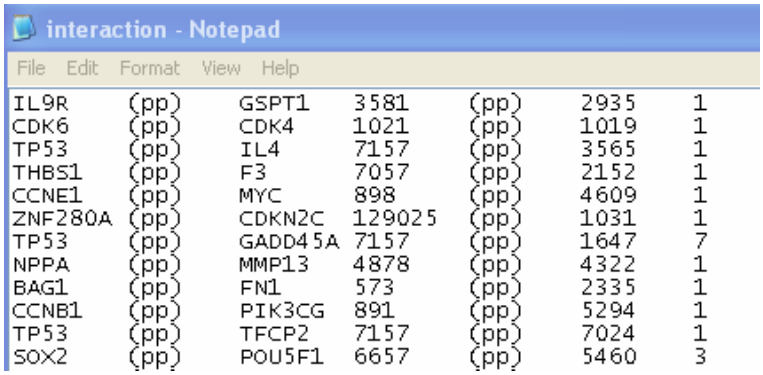
Co-citation protein interaction network is defined as a protein-protein interaction network where two proteins are connected if they are co-cited in one abstract or in the same sentence with one interaction keyword in the text. User is able to query for co-citation network of one gene/protein or many at abstract level or sentence level. We use TP53 as an example:

1. Use TP53 as the keyword to query SCAIView. Select ‘Human Genes/Proteins’ in the Entity tree and tick the ‘Sentence’ box to narrow down the co-citations to the sentence level.

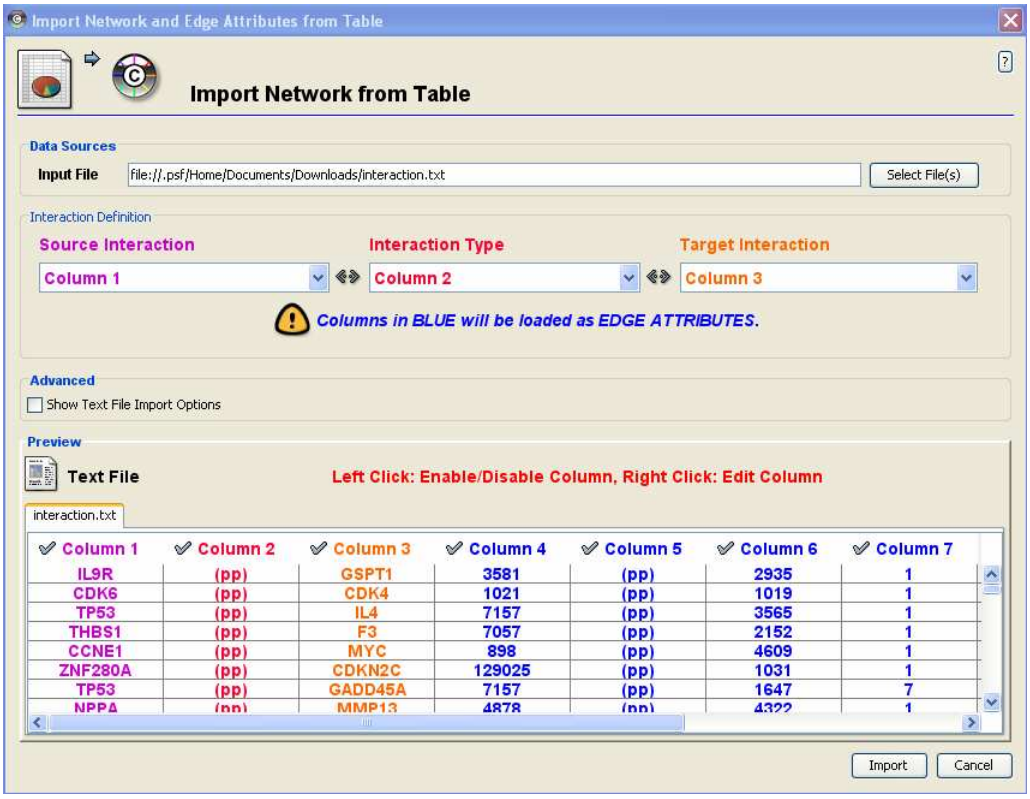


2. Export the results to co-citation network.

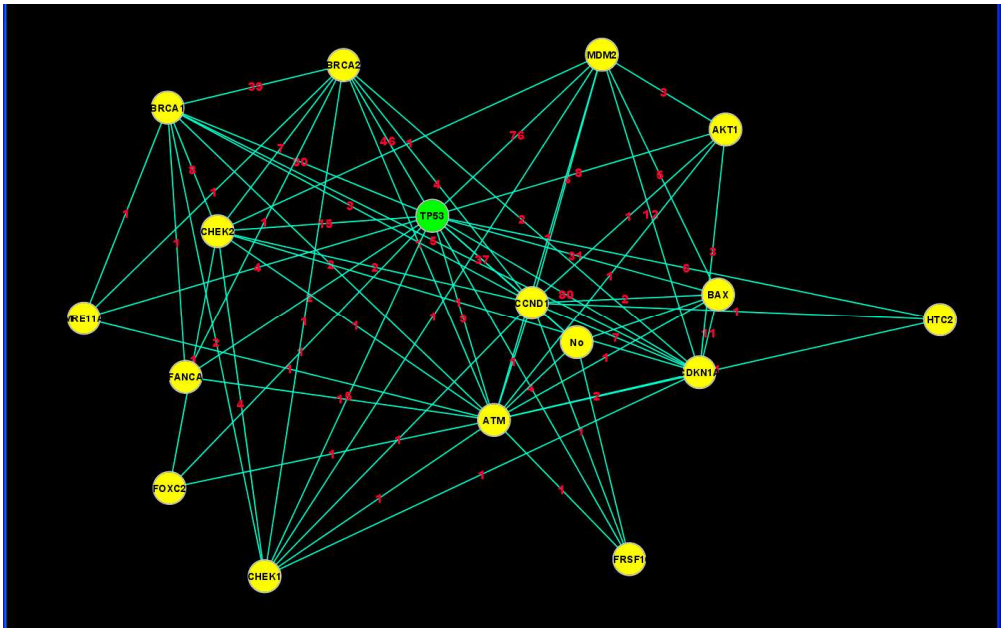
Click on the 'Export PPI' button to export co-citation network in sentence level. The export file is in tab-delimited text format. In export file, column 1 and 3 are interactors while column 2 describes the type of interaction (usually pp or protein-protein interaction); 4th and 6th columns represent the Entrez gene identifiers of the proteins in columns 1 and 2 respectively. Column 7 is the frequency of co-citations.


interaction - Notepad
File Edit Format View Help
IL9R (pp) GSPT1 3581 (pp) 2935 1
CDK6 (pp) CDK4 1021 (pp) 1019 1
TP53 (pp) IL4 7157 (pp) 3565 1
THBS1 (pp) F3 7057 (pp) 2152 1
CCNE1 (pp) MYC 898 (pp) 4609 1
ZNF280A (pp) CDKN2C 129025 (pp) 1031 1
TP53 (pp) GADD45A 7157 (pp) 1647 7
NPPA (pp) MMP13 4878 (pp) 4322 1
BAG1 (pp) FN1 573 (pp) 2335 1
CCNB1 (pp) PIK3CG 891 (pp) 5294 1
TP53 (pp) TFCP2 7157 (pp) 7024 1
SOX2 (pp) POU5F1 6657 (pp) 5460 3

3. Import network into Cytoscape using **File -> import -> import Network from table** in the form of Tab-delimited. Select Column 1 as Source Interaction, column 2 as Interaction Type, column 3 as Target Interaction, and additional columns as edge attributes. Then click **import** button.


Import Network and Edge Attributes from Table
Import Network from Table
Data Sources
Input File
file://.psf/Home/Documents/Downloads/interaction.txt
Select File(s)
Interaction Definition
Source Interaction
Interaction Type
Target Interaction
Column 1
Column 2
Column 3
Columns in BLUE will be loaded as EDGE ATTRIBUTES.
Advanced
Show Text File Import Options
Preview
Text File
Left Click: Enable/Disable Column, Right Click: Edit Column
interaction.txt
Column 1 Column 2 Column 3 Column 4 Column 5 Column 6 Column 7
IL9R (pp) GSPT1 3581 (pp) 2935 1
CDK6 (pp) CDK4 1021 (pp) 1019 1
TP53 (pp) IL4 7157 (pp) 3565 1
THBS1 (pp) F3 7057 (pp) 2152 1
CCNE1 (pp) MYC 898 (pp) 4609 1
ZNF280A (pp) CDKN2C 129025 (pp) 1031 1
TP53 (pp) GADD45A 7157 (pp) 1647 7
Import
Cancel

4. Visualize the co-citation network using Cytoscape: each node represents a protein which is connected to another protein via an edge. The co-occurrence frequencies can be shown as edge labels.



Filtering option:

The system provides users with two filtering functionalities: the strict entity filtering indicated by ‘E’, and the cross-entity document filtering indicated by ‘D’.

1) ‘E’ filtration: Despite “exclusive” types of filter, this is an “inclusive” filter meaning that it limits the context of the search to the terms which are provided by the user. For example, if you're going to extract co-citations of genes involved in for example breast cancer, you can limit your search to those genes or proteins and find those genes that are co-cited with one or several of these specific genes. You can do this either by entering the list of your specific genes/proteins into the filtering field in a space separated format (e.g. list of genes from microarray data) or by adding the genes from the result list directly by pressing the ‘E’ icon in front of each gene/protein).

2) ‘D’ filtration: this option enables the user to filter the retrieved documents for other types of entities which are mentioned together with the current entity. For example, we have queried the system with the keyword ‘breast cancer’ and searched for genes/proteins involved in this disease. Now we would like to know which drugs have been developed for targeting ESR1, the top ranking protein in the list. First we press the ‘D’ filter icon in front of the ESR1 gene and this entity is inserted into the filtering field automatically. Then we select the entity class Drug Names from the tree and repeat the search. The result is a list of drug names which are reported to target ESR1 protein.

A useful feature of filtering option is ‘ontological search’, i.e. in the Entity View several entities of pathways or GO identifiers can be copied into the filter field (through clicking the filter icons in front of each entity) and a new query can be made using these filters; for example a query for the keyword “breast cancer” and searching for the Human Genes/Proteins would result in the list of KEGG pathways annotated to each gene/protein (you must have already added the KEGG pathway column from the Select Columns menu); now clicking on the filter icon in front of one pathway of interest (e.g. Erbb signalling pathway) will insert the identifier of that pathway (e.g. 04012@KEGG) into the filter field; pressing the search button again brings up all the genes annotated to this pathway. The identifier(s) can be also inserted manually in the filter filed in the form of (identifier@entity_type).

Currently, this filtering option can be applied to the following domains: GO identifiers, REACTOME and KEGG pathways, ENTREZGENE identifiers, SWISSPROT identifiers, HUGO gene names, and CHROMOSOME identifiers. It is also possible to apply this

functionality to list the following information in the Entity view: only proteins which are targeted by drugs (1@TARGET), only drugs which are known as biological molecules (1@DRUGBIO), only SNPs which are genotyped by specific array types (1@AFFY100K, 1@AFFY500K, 1@AFFYSNP6, 1@ILLUMINA650Y, 1@ILLUMINA610QUAD, 1@ILLUMINAHUMAN1M). Please note that these results are Boolean, i.e. whether a protein is drug target or not, or whether a drug is biological molecule or not, etc. In contrary to its application to the non-Boolean domains which provides a list of entities (e.g. proteins) that share a certain annotation (e.g. protein domain), this filtering for Boolean results provides a list of all entities (e.g. proteins) which have positive Boolean result (e.g. all are drug targets) without any inference about association between the entities. If inference of association between entities is desired (e.g. proteins that are targets of the same drug), then the same query should be repeated on the entity of interest (e.g. Drug Names only).

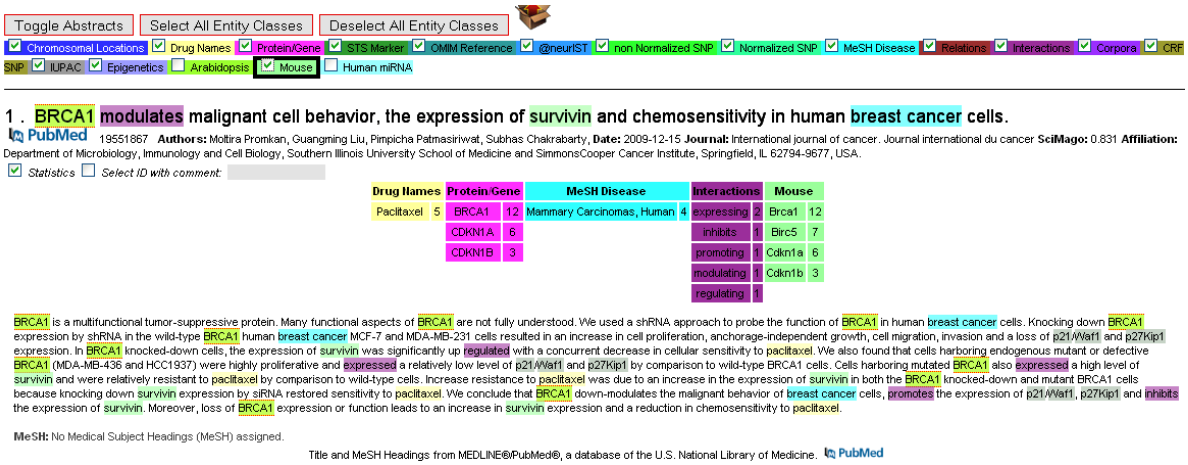
3.4.2 Document View

Document view displays all the documents containing the selected entity from the Entity view and the corresponding search query. By default, these documents are ranked by date (the newest at the top).

On top of the Documents tab, you can select the different entity classes that should be highlighted. When you select an entity type, it is bordered in black. When multiple entities selected, the latest one is bordered. You're also able to select all entity classes at once, deselect all at once, or toggle the abstracts by clicking the corresponding buttons above the colourful entity types.

In the text body, you see your selected entity in yellow which is highlighted differently other than the rest of entities. The colours in the document view are a dimmed version of the colours you find in the legend. In this way overlapping entities are resolved as overlapping colours (they are in general darker). By clicking the checkmark buttons you are able to highlight one of, several, or all the entities found in the abstracts; thus you can focus on a certain entity class or several in combination. When you are focused on one entity class, documents that do not contain elements of this entity class will be greyed out (lighter grey). If you move the pointer over a tagged entity in the text, a tool tip appears, that gives you the possibility to link-out to the specific database providing you additional information on this entity (e.g. EntrezGene). Please note that the tool tip becomes active only for your latest selection of entity type.

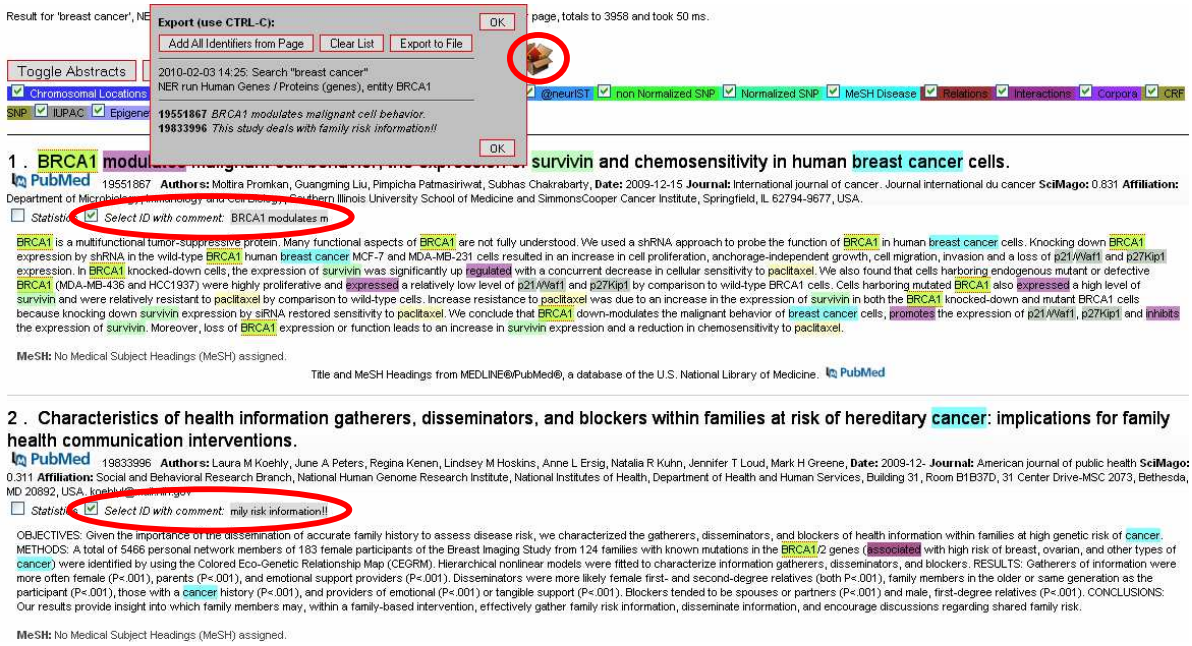
Below the title left to the PubMed ID of each abstract, you find a PubMed Icon that directs users to the abstract of the document at PubMed in an external viewer (providing in some cases access to the full-text of the document). Among the document's information such as authors, date, and journal's name, the "SciMago" index of the journal has been provided, wherever available. SciMago is a type of impact factor based on the page-rank algorithm introduced by an independent Spanish university. Moreover, a "Statistics" option has been also provided in order to give an overall overview of the entities' co-mentionings in the same abstract.


Toggle Abstracts
Select All Entity Classes
Deselect All Entity Classes
Chromosomal Locations
Drug Names
Protein/Gene
STS Marker
OMIM Reference
@neuroIST
non Normalized SNP
Normalized SNP
MeSH Disease
Relations
Interactions
Corpora
CRF
SNP
IUPAC
Epigenetics
Arabidopsis
Mouse
Human miRNA
1 . BRCA1 modulates malignant cell behavior, the expression of survivin and chemosensitivity in human breast cancer cells.
PubMed 19551867 Authors: Moltira Promkan, Guangming Liu, Pimpicha Patmasiriwat, Subhas Chakrabarty, Date: 2009-12-15 Journal: International journal of cancer. Journal international du cancer SciMago: 0.831 Affiliation: Department of Microbiology, Immunology and Cell Biology, Southern Illinois University School of Medicine and SimmonsCooper Cancer Institute, Springfield, IL 62794-9677, USA.
Statistics
Select ID with comment:
Drug Names
Protein/Gene
MeSH Disease
Interactions
Mouse
Paclitaxel 5
BRCA1 12
CDKN1A 6
CDKN1B 3
Mammary Carcinomas, Human 4
expressing 2
inhibits 1
promoting 1
modulating 1
regulating 1
Brca1 12
Birc5 7
Cdkn1a 6
Cdkn1b 3
BRCA1 is a multifunctional tumor-suppressive protein. Many functional aspects of BRCA1 are not fully understood. We used a shRNA approach to probe the function of BRCA1 in human breast cancer cells. Knocking down BRCA1 expression by shRNA in the wild-type BRCA1 human breast cancer MCF-7 and MDA-MB-231 cells resulted in an increase in cell proliferation, anchorage-independent growth, cell migration, invasion and a loss of p21/Waf1 and p27Kip1 expression. In BRCA1 knocked-down cells, the expression of survivin was significantly up regulated with a concurrent decrease in cellular sensitivity to paclitaxel. We also found that cells harboring endogenous mutant or defective BRCA1 (MDA-MB-436 and HCC1937) were highly proliferative and expressed a relatively low level of p21/Waf1 and p27Kip1 by comparison to wild-type BRCA1 cells. Cells harboring mutated BRCA1 also expressed a high level of survivin and were relatively resistant to paclitaxel by comparison to wild-type cells. Increase resistance to paclitaxel was due to an increase in the expression of survivin in both the BRCA1 knocked-down and mutant BRCA1 cells because knocking down survivin expression by siRNA restored sensitivity to paclitaxel. We conclude that BRCA1 down-modulates the malignant behavior of breast cancer cells, promotes the expression of p21/Waf1, p27Kip1 and inhibits the expression of survivin. Moreover, loss of BRCA1 expression or function leads to an increase in survivin expression and a reduction in chemosensitivity to paclitaxel.
MeSH: No Medical Subject Headings (MeSH) assigned.
Title and MeSH Headings from MEDLINE®/PubMed®, a database of the U.S. National Library of Medicine. PubMed

A direct link to the free full-text versions of the documents in PubMed Central (PMC) has been provided when they are available.

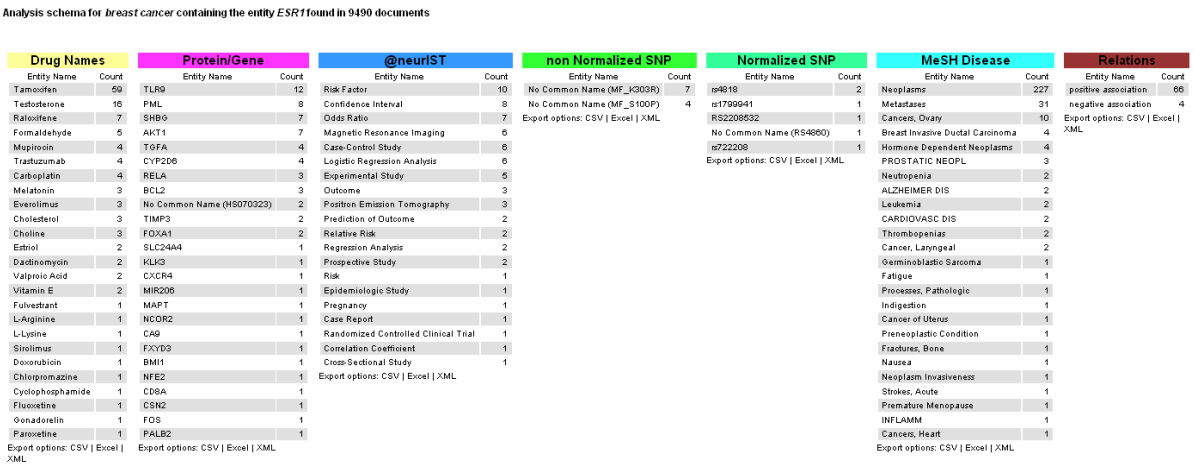
24. Genomewide linkage and haplotype-association studies map intracranial aneurysm to chromosome 7q11.
PubMed
PubMedCentral
11536080 Authors: H Onda, H Kasuya, T Yoneyama, K Takakura, T Hori, J Takeda, T Nakajima, I Inoue, Date: 2001-10- Journal: American journal of human genetics
Statistics

A reporting feature which has been implemented besides the "Statistics" option, allows to export the PubMed ID of the selected abstracts as well as any comments by the user to a text file. It is also possible to copy & paste the sentences of interest from the abstract into the comment field. After the selection is made and/or comments are filled, the abstract ID and comments can be exported to a txt file by pressing the Export to File button. Please note that paging the document view does not affect the selections made in the previous pages.


Export (use CTRL-C):
OK
Add All Identifiers from Page
Clear List
Export to File
2010-02-03 14:25: Search "breast cancer"
NER run Human Genes / Proteins (genes), entity BRCA1
19551867 BRCA1 modulates malignant cell behavior.
19833996 This study deals with family risk information!!
Toggle Abstracts
Chromosomal Locations
@neurIST
non Normalized SNP
Normalized SNP
MeSH Disease
Relations
Interactions
Corpora
CRF
1 . BRCA1 survivin and chemosensitivity in human breast cancer cells.
PubMed 19551867 Authors: Moitira Promkan, Guangming Liu, Pimpicha Patmasiriwat, Subhas Chakrabarty, Date: 2009-12-15 Journal: International journal of cancer. Journal international du cancer SciMago: 0.831 Affiliation:
Select ID with comment: BRCA1 modulates m
BRCA1 is a multifunctional tumor-suppressive protein. Many functional aspects of BRCA1 are not fully understood. We used a shRNA approach to probe the function of BRCA1 in human breast cancer cells. Knocking down BRCA1 expression by shRNA in the wild-type BRCA1 human breast cancer MCF-7 and MDA-MB-231 cells resulted in an increase in cell proliferation, anchorage-independent growth, cell migration, invasion and a loss of p21/Waf1 and p27Kip1 expression. In BRCA1 knocked-down cells, the expression of survivin was significantly up regulated with a concurrent decrease in cellular sensitivity to paclitaxel. We also found that cells harboring endogenous mutant or defective BRCA1 (MDA-MB-436 and HCC1937) were highly proliferative and expressed a relatively low level of p21/Waf1 and p27Kip1 by comparison to wild-type BRCA1 cells. Cells harboring mutated BRCA1 also expressed a high level of survivin and were relatively resistant to paclitaxel by comparison to wild-type cells. Increase resistance to paclitaxel was due to an increase in the expression of survivin in both the BRCA1 knocked-down and mutant BRCA1 cells because knocking down survivin expression by siRNA restored sensitivity to paclitaxel. We conclude that BRCA1 down-modulates the malignant behavior of breast cancer cells, promotes the expression of p21/Waf1, p27Kip1 and inhibits the expression of survivin. Moreover, loss of BRCA1 expression or function leads to an increase in survivin expression and a reduction in chemosensitivity to paclitaxel.
MeSH: No Medical Subject Headings (MeSH) assigned.
Title and MeSH Headings from MEDLINE®/PubMed®, a database of the U.S. National Library of Medicine. PubMed
2 . Characteristics of health information gatherers, disseminators, and blockers within families at risk of hereditary cancer: implications for family health communication interventions.
PubMed 19833996 Authors: Laura M Koehly, June A Peters, Regina Kenen, Lindsey M Hoskins, Anne L Ersig, Natalia R Kuhn, Jennifer T Loud, Mark H Greene, Date: 2009-12- Journal: American journal of public health SciMago: 0.311 Affiliation: Social and Behavioral Research Branch, National Human Genome Research Institute, National Institutes of Health, Department of Health and Human Services, Building 31, Room B1B37D, 31 Center Drive-MSC 2073, Bethesda, MD 20892, USA.
Select ID with comment: mily risk information!!
OBJECTIVES: Given the importance of the dissemination of accurate family history to assess disease risk, we characterized the gatherers, disseminators, and blockers of health information within families at high genetic risk of cancer. METHODS: A total of 5466 personal network members of 183 female participants of the Breast Imaging Study from 124 families with known mutations in the BRCA1/2 genes (associated with high risk of breast, ovarian, and other types of cancer) were identified by using the Colored Eco-Genetic Relationship Map (CEGRM). Hierarchical nonlinear models were fitted to characterize information gatherers, disseminators, and blockers. RESULTS: Gatherers of information were more often female (P<.001), parents (P<.001), and emotional support providers (P<.001). Disseminators were more likely female first- and second-degree relatives (both P<.001), family members in the older or same generation as the participant (P<.001), those with a cancer history (P<.001), and providers of emotional (P<.001) or tangible support (P<.001). Blockers tended to be spouses or partners (P<.001) and male, first-degree relatives (P<.001). CONCLUSIONS: Our results provide insight into which family members may, within a family-based intervention, effectively gather family risk information, disseminate information, and encourage discussions regarding shared family risk.
MeSH: No Medical Subject Headings (MeSH) assigned.

3.4.3 Analysis View

Analysis View provides a statistical overview of the all entities co-mentioned with the entity of interest in the relevant abstracts. By clicking on the analysis icon in front of each entity in the Entity view, the user is directed to the Analysis view page where a statistical overview of co-mentioned genes, GO terms, etc. is shown in separate tables. The statistics of each table can be exported to file with three different formats: CSv, Excel, and XML.



3.5 Application Scenarios

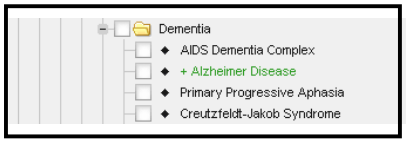
QUERY 1:

Find all the genes that are mentioned in the literature to be associated with the Alzheimer's disease, annotate them with the pathways they're involved in, and keep only those genes which are annotated in the KEGG database to the Alzheimer's disease pathway.

.

SOLUTION:

Step 1) Type in the grey search field "alzheimer".



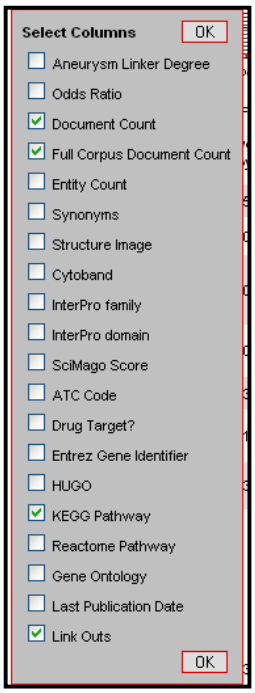
Step 2) Restrict your search to the instances (synonyms) of the Alzheimer's disease in the MeSH disease classification tree under MeSH Disease >> Nervous System Diseases >> Central Nervous System Disease >> Brain Diseases >> Dementia >> Alzheimer's Disease.

HINT: For a correct selection, the entity class name must turn into Green.

Step 3) Select Human Genes/Proteins entity class.

HINT: The position of small magnifier symbol determines which entity must be returned as result.

Step 4) Press search.

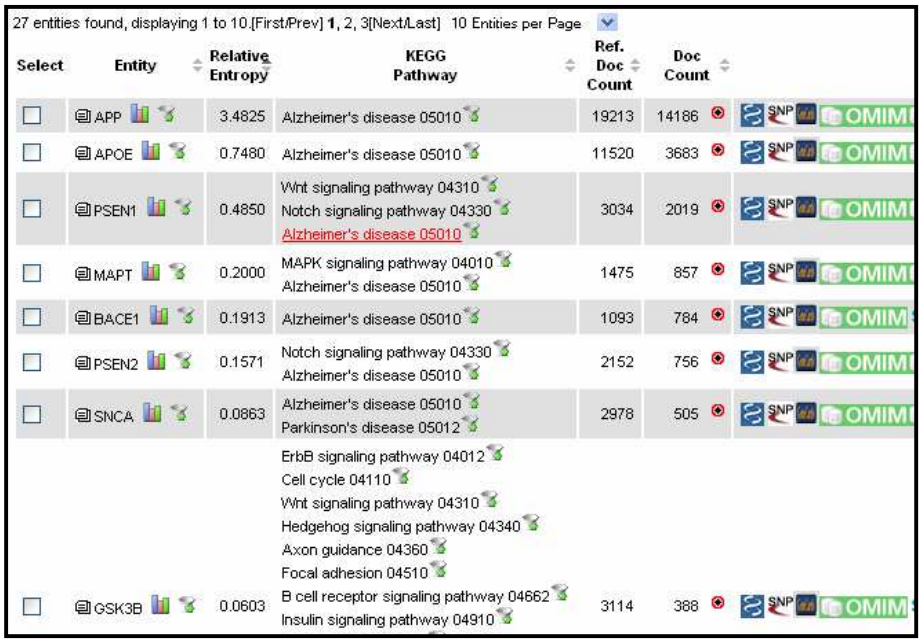


Step 5) Click on the 'Select Columns' button and from the menu select KEGG Pathway, then press OK.

Step 6) In the results table, click on the Filter icon in front of the annotation with *Alzheimer's disease 05010*.

Step 7) Press search again.

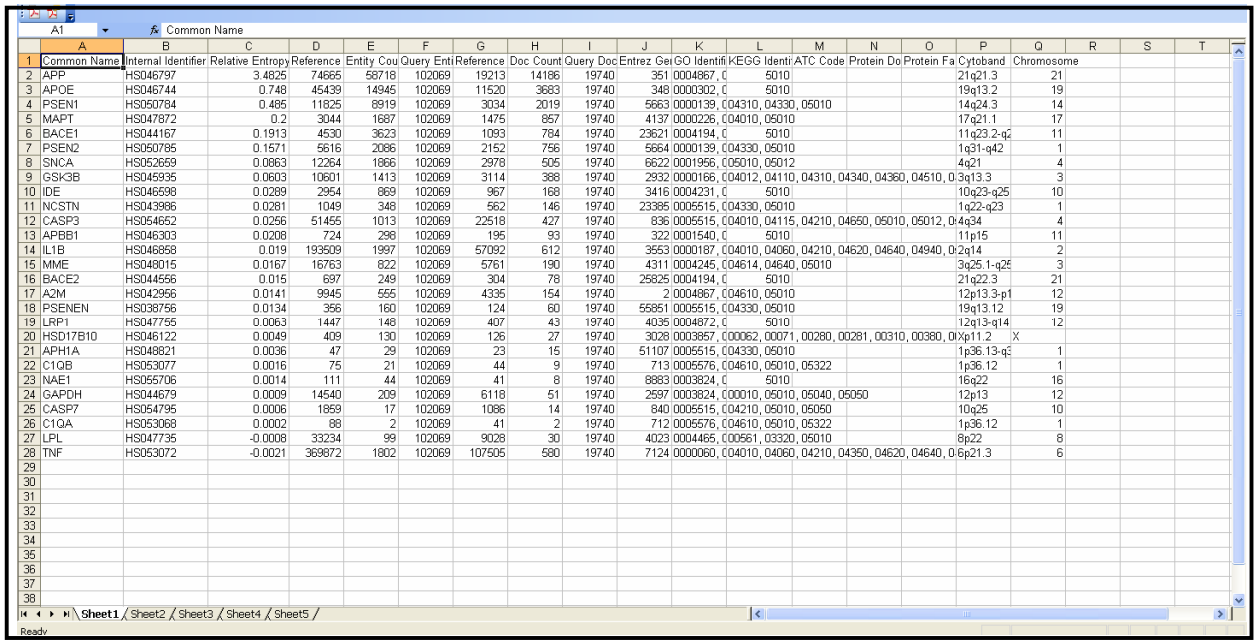
Step 8) Now you should be able to see your results in the Entity View as a list of 27 genes/proteins which are all annotated to the *Alzheimer's disease pathway* in the KEGG database (alongside with other pathways) and ranked according to their relevance to the Alzheimer's disease (see the next page).



27 entities found, displaying 1 to 10.[First/Prev] **1**, 2, 3[Next/Last] 10 Entities per Page

Select	Entity	Relative Entropy	KEGG Pathway	Ref. Doc Count	Doc Count
☐	APP	3.4825	Alzheimer's disease 05010	19213	14186
☐	APOE	0.7480	Alzheimer's disease 05010	11520	3683
☐	PSEN1	0.4850	Wnt signaling pathway 04310 Notch signaling pathway 04330 Alzheimer's disease 05010	3034	2019
☐	MAPT	0.2000	MAPK signaling pathway 04010 Alzheimer's disease 05010	1475	857
☐	BACE1	0.1913	Alzheimer's disease 05010	1093	784
☐	PSEN2	0.1571	Notch signaling pathway 04330 Alzheimer's disease 05010	2152	756
☐	SNCA	0.0863	Alzheimer's disease 05010 Parkinson's disease 05012	2978	505
☐	GSK3B	0.0603	ErbB signaling pathway 04012 Cell cycle 04110 Wnt signaling pathway 04310 Hedgehog signaling pathway 04340 Axon guidance 04360 Focal adhesion 04510 B cell receptor signaling pathway 04662 Insulin signaling pathway 04910	3114	388

Step 9) Click on the Export Table button and save the results in a text file. This file can be imported into an Excel sheet to create your own knowledge base.



Common Name	Internal Identifier	Relative Entropy	Reference	Entity Cou	Query Enti	Reference	Doc Count	Query Doc	Entrez Ge	GO Identifi	KEGG Identi	ATC Code	Protein Do	Protein Fa	Cytoband	Chromosome
APP	HS046797	3.4825	74665	58718	102069	19213	14186	19740	351	0004867,	5010				21q21.3	21
APOE	HS046744	0.748	45439	14945	102069	11520	3683	19740	348	0000302,	5010				19q13.2	19
PSEN1	HS050784	0.485	11825	8919	102069	3034	2019	19740	5663	0000139,	04310, 04330, 05010				14q24.3	14
MAPT	HS047872	0.2	3044	1687	102069	1475	857	19740	4137	0000226,	04010, 05010				17q21.1	17
BACE1	HS044167	0.1913	4530	3623	102069	1093	784	19740	23621	0004194,	5010				11q23.2-q2	11
PSEN2	HS050785	0.1571	5616	2086	102069	2152	756	19740	5664	0000139,	04330, 05010				1q31-q42	1
SNCA	HS052659	0.0863	12264	1866	102069	2978	505	19740	6622	0001956,	05010, 05012				4q21	4
GSK3B	HS045935	0.0603	10601	1413	102069	3114	388	19740	2932	0000166,	04012, 04110, 04310, 04340, 04360, 04510, 0				3q13.3	3
IDE	HS046598	0.0289	2954	869	102069	967	168	19740	3416	0004231,	5010				10q23-q25	10
NCSTN	HS043986	0.0281	1049	348	102069	562	146	19740	23385	0005515,	04330, 05010				1q22-q23	1
CASP3	HS054652	0.0256	51455	1013	102069	22518	427	19740	836	0005515,	04010, 04115, 04210, 04650, 05010, 05012, 0				4q34	4
APBB1	HS046303	0.0208	724	298	102069	195	93	19740	322	0001540,	5010				11p15	11
IL1B	HS046858	0.019	193509	1997	102069	57092	612	19740	3553	0000187,	04010, 04060, 04210, 04620, 04640, 04940, 0				2q14	2
MME	HS048015	0.0167	16763	822	102069	5761	190	19740	4311	0004245,	04614, 04640, 05010				3q25.1-q25	3
BACE2	HS044556	0.015	697	249	102069	304	78	19740	25825	0004194,	5010				21q22.3	21
A2M	HS042956	0.0141	9945	555	102069	4335	154	19740	2	0004867,	04610, 05010				12p13.3-p1	12
PSENEN	HS038756	0.0134	356	160	102069	124	60	19740	55851	0005515,	04330, 05010				19q13.12	19
LRP1	HS047755	0.0063	1447	148	102069	407	43	19740	4035	0004872,	5010				12q13-q14	12
HSD17B10	HS046122	0.0049	409	130	102069	126	27	19740	3028	0003857,	00062, 00071, 00280, 00281, 00310, 00380, 0				Xp11.2	X
APH1A	HS048821	0.0036	47	29	102069	23	15	19740	51107	0005515,	04330, 05010				1p36.13-q3	1
C1QB	HS053077	0.0016	75	21	102069	44	9	19740	713	0005576,	04610, 05010, 05322				1p36.12	1
NAE1	HS055706	0.0014	111	44	102069	41	8	19740	8883	0003824,	5010				16q22	16
GAPDH	HS044679	0.0009	14540	209	102069	6118	51	19740	2597	0003824,	00010, 05010, 05040, 05050				12p13	12
CASP7	HS054795	0.0006	1859	17	102069	1086	14	19740	840	0005515,	04210, 05010, 05050				10q25	10
C1QA	HS053068	0.0002	88	2	102069	41	2	19740	712	0005576,	04610, 05010, 05322				1p36.12	1
LPL	HS047735	-0.0008	33234	99	102069	9028	30	19740	4023	0004465,	00561, 03320, 05010				8p22	8
TNF	HS053072	-0.0021	369872	1802	102069	107505	580	19740	7124	0000060,	04010, 04060, 04210, 04350, 04620, 04640, 0				6p21.3	6

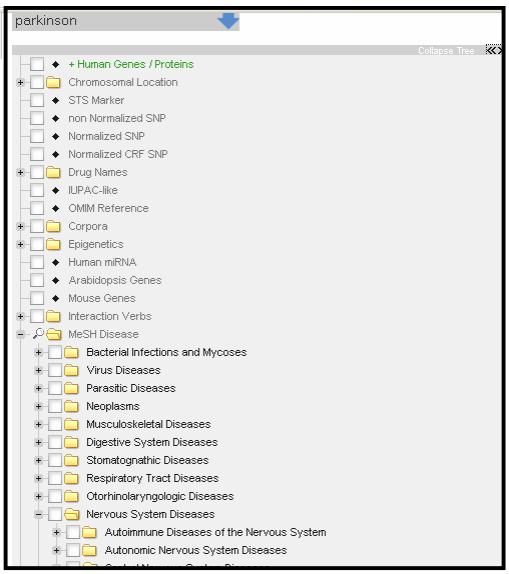
Step 10) Press 'Reset' botton on the left top of the page to start a new search.

QUERY 2:

Find all those proteins that are known drug targets for the Parkinson's disease in the KEGG Parkinson's disease pathway and reconstruct a co-citation network out of them.

SOLUTION:

Step 1) Type in the query field "parkinson".

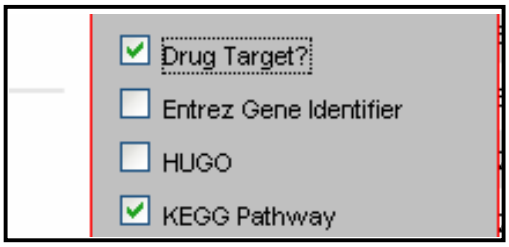


Step 2) Include all possible synonyms of 'Parkinson's disease' by selecting the corresponding entity type in the MeSH classification tree under MeSH Disease >> Nervous System Diseases >> Central Nervous System Disease >> Movement Disorders >> Parkinsonian Disorders >> Parkinson Disease.

Step 3) Select Human Genes/Proteins from entity tree.

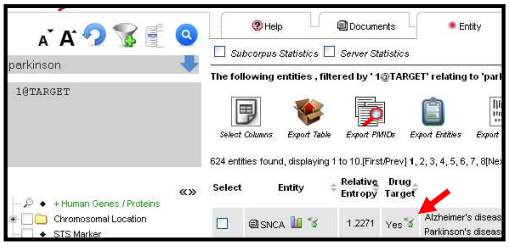
Step 4) Press search.

Step 5) Click on the *Select Columns* button and from the menu select the *KEGG Pathway* and *Drug Target* options, then press OK.



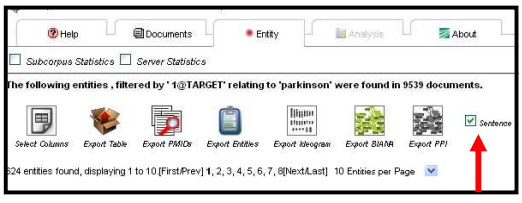
Step 6) Click *Filtering* icon in front of the first target with 'Yes' under the Drug Target column.

Step 7) Press search again.



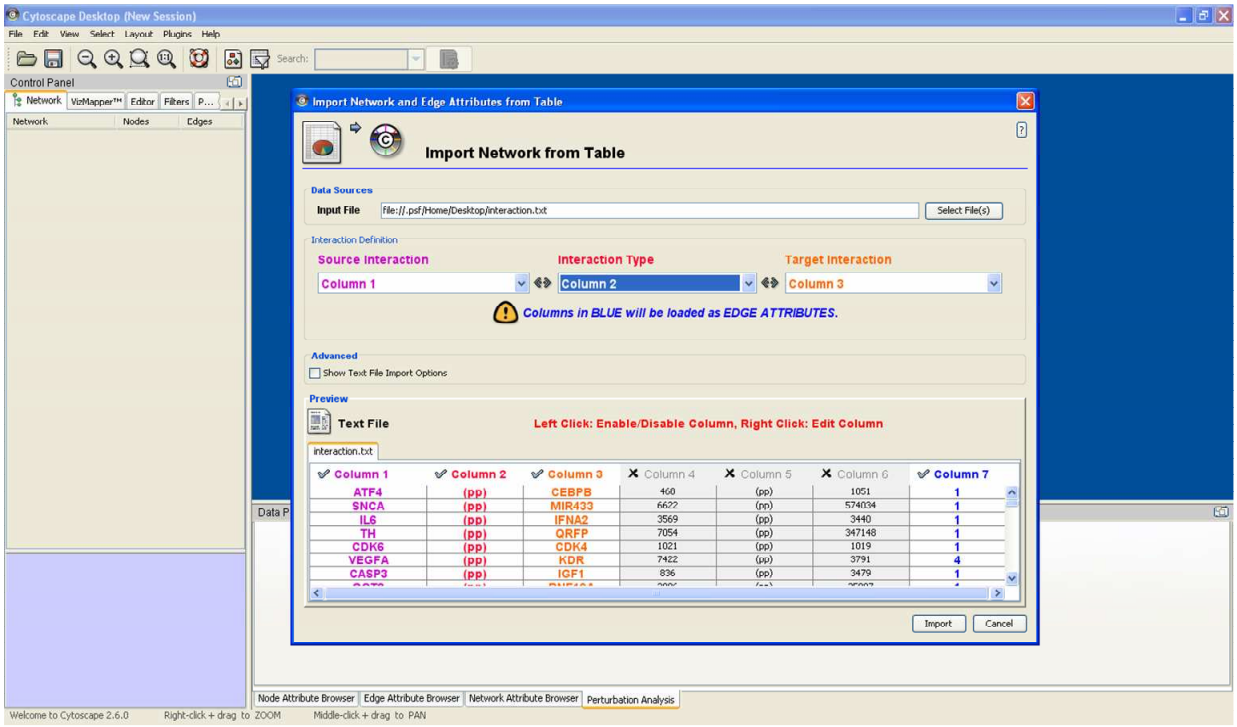
Step 8) A list of genes/proteins which are known drug targets is shown (624 entities found).

Step 9) Now for getting a co-citation network, we restrict the co-citation extraction to the sentence level by ticking the *Sentence* checkbox in front of the *Export PPI* botton.

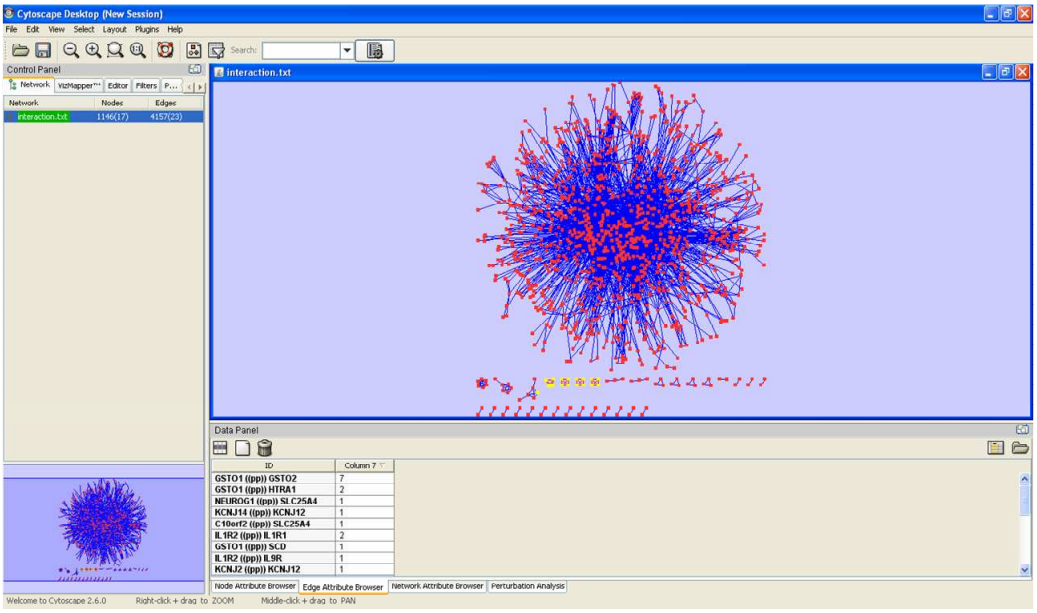


Step 10) Click on the *Export PPI* button to export the co-citation network as a text file which can be imported directly into the Cytoscape environment.

Step 11) Import the file into the Cytoscape environment. Go to Layout → Cytoscape Layouts and click Spring Embedded.



Step 12) The target co-citation network for Parkinson's disease is ready for further analysis in the Cytoscape environment.



QUERY 3:

What SNPs (single nucleotide polymorphisms) are targeted by drugs in the Epilepsy disease? Report their distribution on the human karyogram and the list of drugs that target them.

SOLUTION:

Step 1) Press 'Reset' botton to start a new search.

Step 2) Type in the query field "epilepsy".

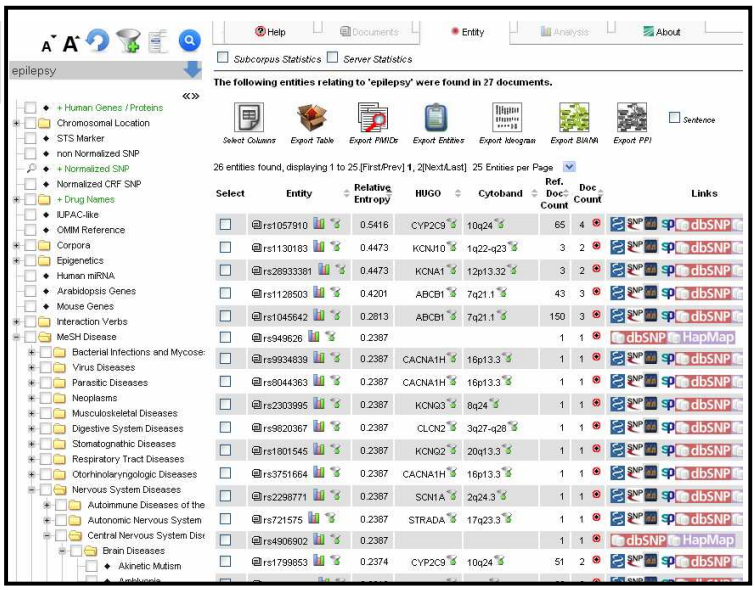
Step 3) Select Epilepsy in the MeSH classification tree under MeSH Disease >> Nervous System Diseases >> Central Nervous System Disease >> Brain Diseases >> Epilepsy.

Step 4) From the Entity tree, select the followings **orderly**:

a) Human Genes/Proteins (to consider those abstracts that contain the name of genes carrying the SNP),

b) Drug Names (to consider those documents which contain drug names related to epilepsy),

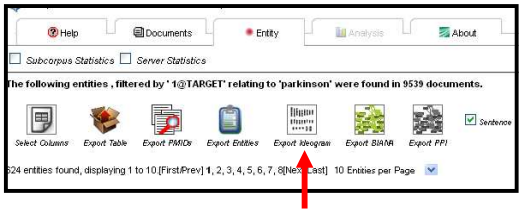
c) Normalized SNP (to retrieve a list of SNPs relevant to epilepsy).


epilepsy
The following entities relating to 'epilepsy' were found in 27 documents.
26 entities found, displaying 1 to 25.[First/Prev] 1, 2[Next/Last] 25 Entities per Page

Step 5) Press search. This should result in a table containing 26 entities.

Step 6) From the *Select Columns* menu, add the *cytoband* and *HUGO* columns. The former lists the chromosomal locations of the SNPs and the latter lists the name of genes corresponding to the SNPs.

Step 7) Now click on the IdeogramBrowser icon and save the SNP cytoband markers in a text file. This file can be loaded into the IdeogramBrowser software environment and SNP cytobands can be visualized over the human karyogram.



Step 8) Open the IdeogramBrowser environment and under File → Load Markers import the text file. You are able now to see the location of SNP makers on the human karyogram.

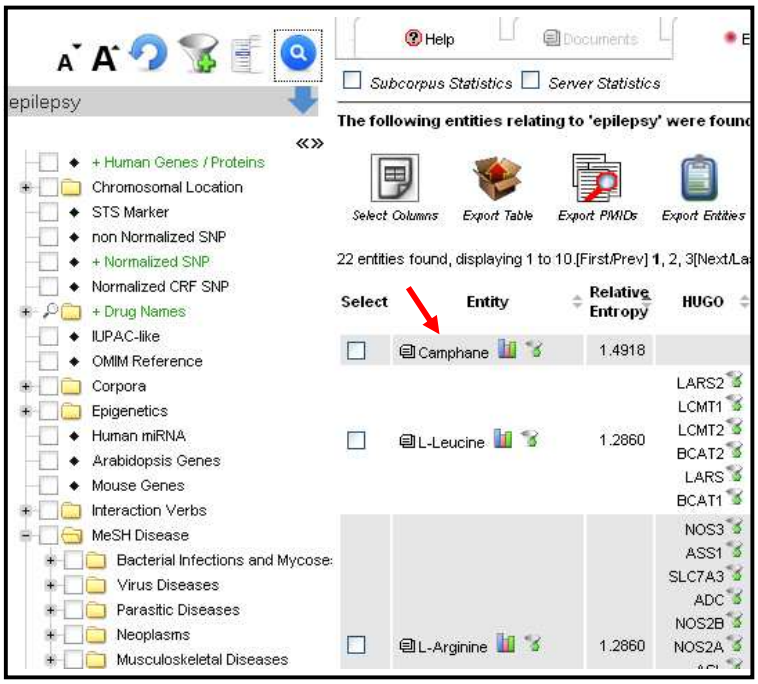
Step 9) Back to the Entity View, move the 'magnifier' symbol from the Normalized SNPs to Drug Names without clicking on them.

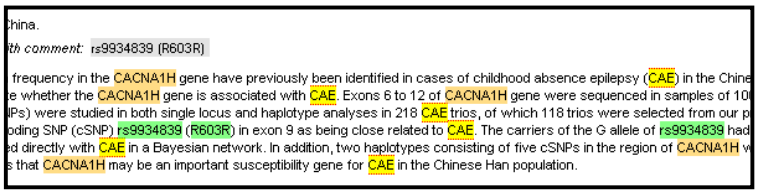
Step 10) Press search

Step 11) Under HUGO column, a list of SNP-carrier genes which are targeted by those drugs is found (if exists). Click on the DrugBank icon to see the properties of the drug in an overlaid page.

Step 12) Click on the first drug name (Camphane) in the Entity table to be directed to the Documents View.

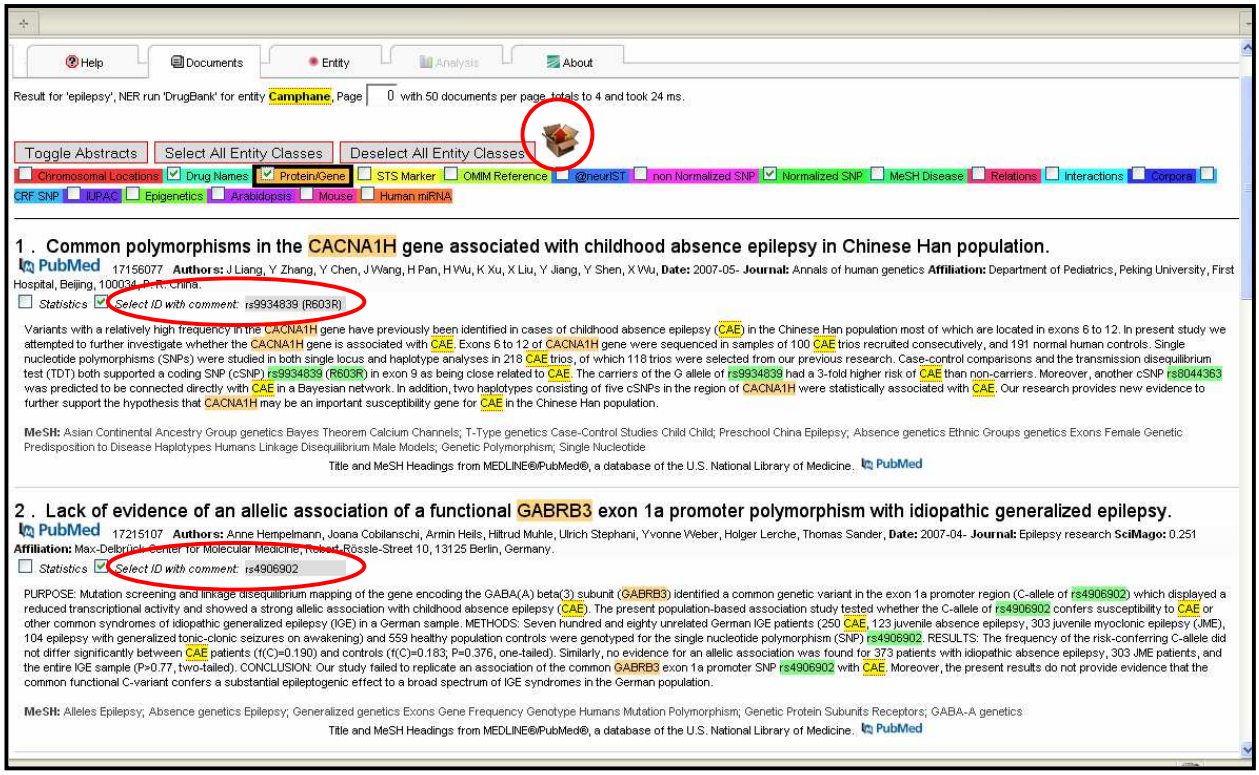
Step 13) To find out which SNPs are targeted by this drug, tick the *Normalized SNP* in the highlight bar. Now you are able to see the corresponding SNPs highlighted in the abstract text.




hina.
th comment: rs9934839 (R603R)
frequency in the CACNA1H gene have previously been identified in cases of childhood absence epilepsy (CAE) in the Chine
e whether the CACNA1H gene is associated with CAE. Exons 6 to 12 of CACNA1H gene were sequenced in samples of 10
IPs) were studied in both single locus and haplotype analyses in 218 CAE trios, of which 118 trios were selected from our p
oding SNP (cSNP) rs9934839 (R603R) in exon 9 as being close related to CAE. The carriers of the G allele of rs9934839 had
d directly with CAE in a Bayesian network. In addition, two haplotypes consisting of five cSNPs in the region of CACNA1H v
s that CACNA1H may be an important susceptibility gene for CAE in the Chinese Han population.

Step 14) Type or copy/paste the name of these SNPs from each abstract into the *Select ID with comment* filed which lies under the title of each abstract. Then tick the checkbox.

Step 15) Click on the Export icon on top of the page to export your comments along with their corresponding PubMed identifiers to a text file.


Help
Documents
Entity
Analysis
About
Result for 'epilepsy', NER run 'DrugBank' for entity Camphane, Page 0 with 50 documents per page. totals to 4 and took 24 ms.
Toggle Abstracts
Select All Entity Classes
Deselect All Entity Classes
Chromosomal Locations
Drug Names
Protein/Gene
STS Marker
OMIM Reference
t2neuriST
non Normalized SNP
Normalized SNP
MeSH Disease
Relations
Interactions
Corpora
CRF SNP
IUPAC
Epigenetics
Arabidopsis
Mouse
Human miRNA
1 . Common polymorphisms in the CACNA1H gene associated with childhood absence epilepsy in Chinese Han population.
PubMed 17156077 Authors: J Liang, Y Zhang, Y Chen, J Wang, H Pan, H Wu, K Xu, X Liu, Y Jiang, Y Shen, X Wu, Date: 2007-05- Journal: Annals of human genetics Affiliation: Department of Pediatrics, Peking University, First Hospital, Beijing, 100034, P.R. China.
Statistics
Select ID with comment: rs9934839 (R603R)
Variants with a relatively high frequency in the CACNA1H gene have previously been identified in cases of childhood absence epilepsy (CAE) in the Chinese Han population most of which are located in exons 6 to 12. In present study we attempted to further investigate whether the CACNA1H gene is associated with CAE. Exons 6 to 12 of CACNA1H gene were sequenced in samples of 100 CAE trios recruited consecutively, and 191 normal human controls. Single nucleotide polymorphisms (SNPs) were studied in both single locus and haplotype analyses in 218 CAE trios, of which 118 trios were selected from our previous research. Case-control comparisons and the transmission disequilibrium test (TDT) both supported a coding SNP (cSNP) rs9934839 (R603R) in exon 9 as being close related to CAE. The carriers of the G allele of rs9934839 had a 3-fold higher risk of CAE than non-carriers. Moreover, another cSNP rs8044363 was predicted to be connected directly with CAE in a Bayesian network. In addition, two haplotypes consisting of five cSNPs in the region of CACNA1H were statistically associated with CAE. Our research provides new evidence to further support the hypothesis that CACNA1H may be an important susceptibility gene for CAE in the Chinese Han population.
MeSH: Asian Continental Ancestry Group genetics Bayes Theorem Calcium Channels; T-Type genetics Case-Control Studies Child Child; Preschool China Epilepsy; Absence genetics Ethnic Groups genetics Exons Female Genetic Predisposition to Disease Haplotypes Humans Linkage Disequilibrium Male Models; Genetic Polymorphism; Single Nucleotide
Title and MeSH Headings from MEDLINE®/PubMed®, a database of the U.S. National Library of Medicine. PubMed
2 . Lack of evidence of an allelic association of a functional GABRB3 exon 1a promoter polymorphism with idiopathic generalized epilepsy.
PubMed 17215107 Authors: Anne Hempelmann, Joana Cobilanschi, Armin Heils, Hiltrud Muhle, Ulrich Stephani, Yvonne Weber, Holger Lerche, Thomas Sander, Date: 2007-04- Journal: Epilepsy research SciMago: 0.251
Affiliation: Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Street 10, 13125 Berlin, Germany.
Statistics
Select ID with comment: rs4906902
PURPOSE: Mutation screening and linkage disequilibrium mapping of the gene encoding the GABA(A) beta(3) subunit (GABRB3) identified a common genetic variant in the exon 1a promoter region (C-allele of rs4906902) which displayed a reduced transcriptional activity and showed a strong allelic association with childhood absence epilepsy (CAE). The present population-based association study tested whether the C-allele of rs4906902 confers susceptibility to CAE or other common syndromes of idiopathic generalized epilepsy (IGE) in a German sample. METHODS: Seven hundred and eighty unrelated German IGE patients (250 CAE, 123 juvenile absence epilepsy, 303 juvenile myoclonic epilepsy (JME), 104 epilepsy with generalized tonic-clonic seizures on awakening) and 559 healthy population controls were genotyped for the single nucleotide polymorphism (SNP) rs4906902. RESULTS: The frequency of the risk-conferring C-allele did not differ significantly between CAE patients (f(C)=0.190) and controls (f(C)=0.183; P=0.376, one-tailed). Similarly, no evidence for an allelic association was found for 373 patients with idiopathic absence epilepsy, 303 JME patients, and the entire IGE sample (P>0.77, two-tailed). CONCLUSION: Our study failed to replicate an association of the common GABRB3 exon 1a promoter SNP rs4906902 with CAE. Moreover, the present results do not provide evidence that the common functional C-variant confers a substantial epileptogenic effect to a broad spectrum of IGE syndromes in the German population.
MeSH: Alleles Epilepsy; Absence genetics Epilepsy; Generalized genetics Exons Gene Frequency Genotype Humans Mutation Polymorphism; Genetic Protein Subunits Receptors; GABA-A genetics
Title and MeSH Headings from MEDLINE®/PubMed®, a database of the U.S. National Library of Medicine. PubMed

Step 16) Press Reset to begin a new search.

Specialized versions of SCAIView for other domains can be developed. Project ideas are welcome. For commercial use, please contact
Dr. Christoph M. Friedrich, friedrich at scai.fraunhofer.de

Publishing Notes - Impressum

The Fraunhofer Institute for Algorithms
and Scientific Computing SCAI
Schloss Birlinghoven
53754 Sankt Augustin
Germany
Phone +49 2241 14-2500
Fax +49 2241 14-2460
http://www.scai.fraunhofer.de
is a constituent entity of the Fraunhofer-Gesellschaft, and as such has no separate legal status.